MEDRESEARCH: A SCREENING-GUIDED AGENT FOR MEDICAL DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical deep research collects evidence that meets predefined eligibility criteria and synthesizes it into reports to support medical decision-making. A well-supported report, however, does not ensure comprehensive coverage of eligible evidence. Task-trained methods require additional supervision, while training-free methods reuse existing models but often use screening mainly to select sources rather than guide subsequent actions. Acting on screening judgments poses two challenges: exclusions may be incorrect, and the amount of undiscovered eligible evidence is unknown. To face these challenges, we introduce **MedResearch**, a training-free agent built around **Screen-to-Steer**: screening judgments guide both evidence retention and subsequent agent actions. Its Protocol Reader reassesses first-pass exclusions with additional text from the same source; Eligibility Memory preserves the resulting judgments and search progress; and the Search Controller uses this state to direct further retrieval, trigger one-shot recovery, or stop. The retained evidence is then synthesized into a structured research report, with all model weights kept frozen throughout execution. Experiments on MetaSyn and CLEF show higher evidence-selection recall and F1 than the evaluated RAG and deep research baselines, together with strong report-level conclusion alignment. Ablations and evidence-flow analysis further highlight the complementary roles of reassessment, memory, and action control in collecting and retaining evidence for synthesis. Code is available in an anonymous repository.[1]

1 INTRODUCTION

Medical deep research retrieves and synthesizes scientific evidence to support clinical decisions and guide medical research (Page et al., 2021; Wang et al., 2025a). Given a medical question, predefined eligibility criteria, and a searchable literature corpus, the task is to collect and retain evidence that meets these criteria as comprehensively as possible, then synthesize it into a research report (Rethlefsen et al., 2021). Automating this workflow draws on the multi-step search and synthesis capabilities of general-purpose deep research (Shao et al., 2024; Du et al., 2026). *However, a well-supported report derived from general-purpose deep research does not imply comprehensive coverage of eligible evidence*. Evidence may be missed during retrieval or mistakenly excluded during screening (Xie et al., 2026). In medicine, omitting evidence of limited benefit or adverse effects can lead to overstated treatment benefits or underestimated harms, potentially misleading clinical decisions (Eyding et al., 2010).

Existing medical deep research approaches can be broadly grouped by whether they require additional task-specific training, as shown in Figure 1. **(i) Task-trained methods**. This thread of methods, *e.g.,* LEADS (Wang et al., 2025b), MedCPT (Jin et al., 2023), and Bio-SIEVE (Robinson et al., 2023), uses additional task-specific training to acquire specialized capabilities for medical evidence retrieval and screening. However, such approaches rely on substantial task-specific data and training, making them costly to build and less convenient to adapt when the target domain or review workflow changes. **(ii) Training-free methods**. By contrast, methods such as TrialMind (Wang et al., 2025a) reuse existing language models and retrieval tools, allowing new medical questions and eligibility criteria to be supplied without updating model parameters. Recent agentic workflows such as ProtoMA (Xie et al., 2026) and MedSR-Copilot (Huang et al., 2026) further coordinate

[1] https://anonymous.4open.science/r/MedResearch_ICLR-588B

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

retrieval, screening, and synthesis. Yet existing agentic workflows use screening, which checks evidence against eligibility criteria, mainly to decide what evidence to keep, rather than to guide what the agent should do next, limiting how evidence collection adapts to what has already been learned.

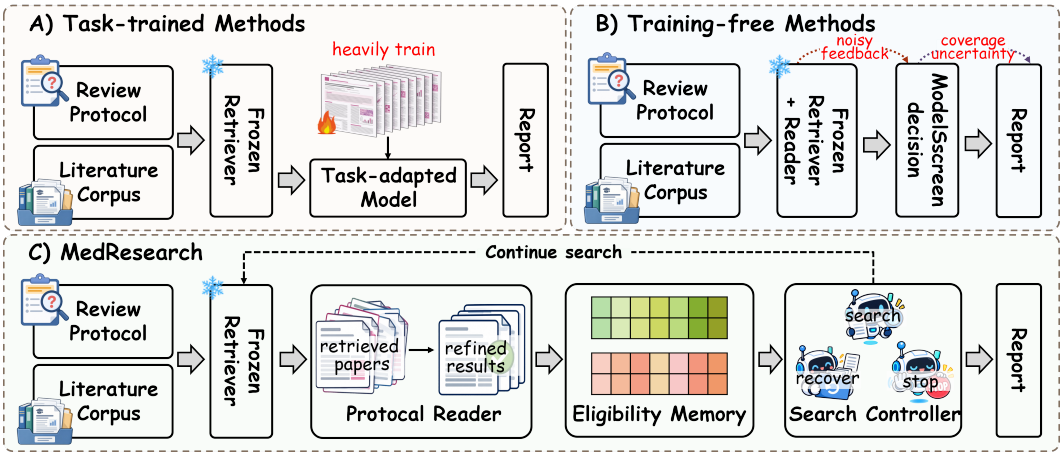


Figure 1: **Two routes to medical deep research.** MedResearch follows the training-free route.

The limitations of training-free workflows motivate **Screen-to-Steer**: screening judgments should guide both what evidence to keep and what the agent does next. For a training-free medical deep research agent, this raises two challenges. **(i) Noisy Feedback**. Limited evidence can cause the agent to mistake missing information for a failed eligibility criterion, so an incorrect exclusion may both discard eligible evidence and misguide later actions (Khraisha et al., 2024; Matsui et al., 2024). **(ii) Coverage Uncertainty**. The agent only observes the evidence collected so far, not the full set of eligible evidence (Callaghan & Müller-Hansen, 2020). Thus, finding no new eligible evidence does not imply that collection is complete: stopping may miss evidence, while continuing incurs additional cost (Boetje & van de Schoot, 2024; Callaghan et al., 2024).

We implement **Screen-to-Steer** as **MedResearch**, a training-free medical deep research agent built on a frozen retriever and language model. To mitigate **Noisy Feedback**, its **Protocol Reader** screens retrieved evidence and reassesses first-pass exclusions with additional text from the same source when available. **Eligibility Memory** preserves the final screening judgments and search progress so they can guide subsequent actions. To manage **Coverage Uncertainty**, the devised **Search Controller** uses the screening history and number of new inclusions to direct retrieval and choose continued search, one-shot recovery, or stopping. During ordinary search, new candidates return to the reader, and their screening results update the memory before the next action. At termination, the agent then generates the retained evidence into a structured research report and returns the supporting sources and an execution trace. Our contributions are threefold:

- We introduce **Screen-to-Steer** for medical deep research, where screening judgments guide both evidence selection and subsequent agent actions, coupling exclusion reassessment with adaptive evidence collection.
- We instantiate this design as **MedResearch**, a training-free medical deep research agent that integrates a **Protocol Reader**, **Eligibility Memory**, and **Search Controller** to connect evidence screening with adaptive evidence collection and report synthesis.
- We evaluate MedResearch on **MetaSyn** and **CLEF**, where it achieves higher **evidence-selection recall and F1** than the evaluated general-purpose and medical deep research methods, and fixed RAG baselines. Report-level evaluation and component ablations further examine synthesis quality and the contributions of the proposed design.

2 Problem Setup

We study **medical deep research**, where a system collects scientific evidence under explicit medical eligibility criteria and synthesizes the selected evidence into a research report. Compared with

general-purpose deep research, this setting makes **which evidence is eligible and how comprehensively it is collected** explicit parts of the task, rather than focusing only on the final report.

Formally, let p denote a medical research protocol and $\mathcal{D}$ a searchable literature corpus. The protocol specifies a medical question together with its inclusion and exclusion criteria. Given $(p, \mathcal{D})$, the system retrieves a candidate set $\mathcal{C}(p) \subseteq \mathcal{D}$ and screens these candidates to produce a predicted eligible-evidence set $\widehat{\mathcal{G}}(p)$:

$$\widehat{\mathcal{G}}(p) \subseteq \mathcal{C}(p) \subseteq \mathcal{D}. \tag{1}$$

Each paper in $\mathcal{D}$ is treated as one evidence record. The goal of evidence collection is to retain eligible records as comprehensively as possible while avoiding incorrect inclusions.

The system then synthesizes the selected evidence into a research report:$\widehat{\mathcal{R}}(p)$:

$$\widehat{\mathcal{R}}(p) = \text{Synthesize}\big(p, \widehat{\mathcal{G}}(p)\big). \tag{2}$$

Thus, medical deep research comprises two connected stages: **evidence collection**, which determines the evidence base, and **report synthesis**, which organizes that evidence into the final research output. Evidence collection may operate on a fixed candidate pool or acquire additional candidates iteratively as research proceeds.

3 MEDRESEARCH

Given a medical research protocol p and a searchable literature corpus $\mathcal{D}$, **MedResearch** implements **Screen-to-Steer** with a frozen retriever and language model. The **Protocol Reader** first determines which retrieved evidence should be retained. Because limited text can lead to premature exclusions, it selectively revisits excluded candidates with additional text from the same source when available. The agent preserves these judgments so they can guide subsequent evidence collection. The **Eligibility Memory** records the reader's final inclusion and exclusion judgments and tracks the progress of evidence collection. The **Search Controller** uses this accumulated information to decide what the agent should do next. Further search returns new candidates to the Protocol Reader, whose judgments update the memory before the next action is taken. Together, these components realize **Screen-to-Steer**: screening judgments determine both what evidence is retained and how subsequent evidence collection proceeds. At termination, the accumulated evidence forms $\widehat{\mathcal{G}}(p)$ and is passed to report synthesis. Figure 2 summarizes the workflow.

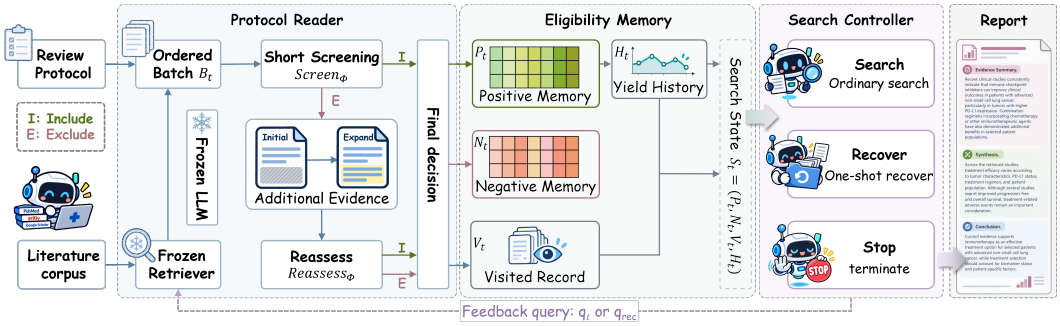


Figure 2: **Overview of MedResearch.** The Protocol Reader produces screening judgments, Eligibility Memory stores them with search progress, and the Search Controller uses this state to select the next action. Ordinary search returns new candidates to the reader, closing the loop. I and E denote predicted INCLUDE and EXCLUDE labels, respectively.

3.1 PROTOCOL READER

Screening from limited text can produce unreliable exclusions when information needed by the eligibility criteria has not yet been observed. The **Protocol Reader** therefore refines screening judgments before they guide later agent actions. It first performs a short screening pass and then selectively reassesses first-pass exclusions using additional text from the same source. In detail, the reader is implemented by a frozen language model L_ϕ, whose parameters ϕ remain fixed throughout

execution. We denote the two operations by Screen_ϕ and $\mathrm{Reassess}_\phi$; both use the same underlying model with distinct screening and reassessment instructions.

Protocol-Guided Screening. Let B denote an ordered batch of candidate records passed to the reader, and let $d \in B$ denote an individual candidate. For each d, the reader constructs an initial reading context x_d^s from its available abstract text and metadata. The superscript s denotes the initial screening pass. Given the review protocol p, the reader produces one label $\widehat{y}_d^s \in \{\mathsf{I}, \mathsf{E}\}$ per candidate:

$$\{\widehat{y}_d^s\}_{d\in B} = \mathrm{Screen}_\phi\big(p, \{x_d^s\}_{d\in B}\big). \tag{3}$$

Here, I (INCLUDE) marks a candidate for retention as evidence, whereas E (EXCLUDE) marks it for exclusion based on the current reading context. First-pass exclusions are revisited before the final screening decision.

Exclusion Reassessment. To reduce premature exclusions, the reader revisits candidates initially labeled EXCLUDE. These candidates form $B_E = \{d \in B : \widehat{y}_d^s = \mathsf{E}\}$, retaining their order in B. For each $d \in B_E$, the reader constructs the reassessment context x_d^l from the initial context and any further available abstract text. The superscript l denotes the expanded-context reassessment pass. The same frozen model produces the reassessed labels:

$$\{\widehat{y}_d^l\}_{d\in B_E} = \mathrm{Reassess}_\phi\big(p, \{x_d^l\}_{d\in B_E}\big). \tag{4}$$

Initial inclusions remain unchanged, while initial exclusions take their reassessed labels. Writing $\widehat{y}_d$ without a superscript for the final reader decision, we have:

$$\widehat{y}_d = \begin{cases} \mathsf{I}, & \widehat{y}_d^s = \mathsf{I}, \\ \widehat{y}_d^l, & \widehat{y}_d^s = \mathsf{E}. \end{cases} \tag{5}$$

An $\mathsf{E} \to \mathsf{I}$ transition withdraws the initial exclusion, whereas $\mathsf{E} \to \mathsf{E}$ upholds it. The final labels are then used for record selection and search-memory updates.

3.2 ELIGIBILITY MEMORY

Eligibility Memory connects the Protocol Reader's final screening judgments to subsequent agent actions. It preserves positive and negative eligibility feedback in query memories P_t and N_t, containing records finally judged INCLUDE and EXCLUDE, respectively. Together with search progress, these memories summarize what the agent has learned from previously screened evidence and provide the state used by the **Search Controller** to guide subsequent evidence collection.

Memory Initialization. The initial batch B_0 is retrieved using the protocol-derived query, ordered by retrieval rank, and processed by the **Protocol Reader**. The selected evidence is retained for final output, while a fixed top-ranked subset of B_0 is used to seed query memory. Within this subset, retained inclusions populate P_0, and records with final EXCLUDE labels populate N_0. All initial candidates are marked visited, $V_0 = B_0$. Detailed initial-selection and memory-seeding rules are provided in **Appendix B.4**.

Memory Update. For each ordinary search round $t \geq 1$, the new candidate batch B_t first passes through the **Protocol Reader**. For each record $d \in B_t$, let $\widehat{y}_d$ denote the reader's final label after screening and any required reassessment. New INCLUDE records are retained for final output and added to positive memory P_t, while records that remain EXCLUDE are added to negative memory N_t. Existing memory entries are retained across rounds. An exclusion reversed by reassessment therefore contributes positive rather than negative feedback. In this way, each screening round updates the state that will guide the agent's subsequent actions.

Search Progress. Alongside screening judgments, the memory tracks what has already been examined and how productive recent search has been. The visited-record set V_t starts with $V_0 = B_0$ and grows with each new batch, allowing retrieval to skip previously screened records. For an ordinary batch B_t, the observed yield is:

$$Y_t = \left|\{d \in B_t : \widehat{y}_d = \mathsf{I}\}\right|, \tag{6}$$

where I denotes INCLUDE. The yield history is $H_t = (Y_1, \ldots, Y_t)$, with H_0 empty.

Together, the query memories, visited records, and yield history form the search state:

$$S_t = (P_t, N_t, V_t, H_t). \tag{7}$$

This state links memory update to subsequent search: P_t and N_t provide eligibility feedback for query construction, V_t prevents repeated retrieval, and H_t summarizes recent search progress. The **Search Controller** then uses S_t to determine the agent's next action.

3.3 SEARCH CONTROLLER

The **Protocol Reader** produces final screening judgments, and **Eligibility Memory** combines them with search progress to update the state S_t. The **Search Controller** then applies a fixed, state-dependent policy to choose SEARCH, one-shot RECOVER, or STOP. After the initial candidates have been screened and memory has established S_0, the controller starts the first ordinary search if unseen candidates remain.

Yield-Guided Actions. After ordinary round t, the controller uses the updated state S_t to select the next action. Here, Y_t counts the new predicted inclusions in that round, and $T_{\max}$ is the maximum number of ordinary search rounds. If no unseen candidates remain, the controller selects STOP. Otherwise, the controller applies the following yield-guided policy:

$$a_t = \begin{cases} \textsc{Recover}, & Y_t = 0,\ t > 1,\ P_t \neq \varnothing, \\ \textsc{Search}, & Y_t > 0,\ t < T_{\max}, \\ \textsc{Stop}, & \text{otherwise}. \end{cases} \tag{8}$$

A zero yield in the first ordinary round stops collection immediately. A later zero yield can trigger recovery when positive memory is available, including at round $T_{\max}$. A productive round continues search unless the round limit has been reached, in which case the agent stops without recovery. Observed yield measures search progress, not whether all eligible evidence has been found.

Memory-Guided Search. To execute ordinary search round t, the controller constructs a query using the negative memory in the preceding state S_{t-1}. Let q_0 be the unit-normalized base query vector derived from the protocol, and $e(d)$ the retriever's unit-normalized document vector. For a nonempty memory F, let $\mu(F) = |F|^{-1} \sum_{d \in F} e(d)$, with $\mu(\varnothing) = 0$. The updated query and candidate batch are then obtained as follows:

$$q_t = \mathrm{norm}(q_0 - \lambda \mu(N_{t-1}))\,, \qquad B_t = \mathrm{Retrieve}(q_t, \mathcal{D} \setminus V_{t-1}, b)\,, \tag{9}$$

where $\mathrm{norm}(z) = z/\|z\|_2$, λ controls feedback strength, and b is the maximum batch size. The update follows vector-space relevance feedback (Salton & Buckley, 1990), shifting the base query away from the average representation of records judged ineligible. Positive memory contributes to query construction only during recovery. The **Protocol Reader** processes the retrieved batch, and **Eligibility Memory** incorporates its final judgments and yield to form S_t, which the controller uses to select the next action.

One-Shot Recovery. When RECOVER is selected after round t, the controller constructs a recovery query using both positive and negative memory:

$$q_{\mathrm{rec}} = \mathrm{norm}(q_0 + \lambda \mu(P_t) - \lambda \mu(N_t))\,. \tag{10}$$

The retriever returns at most b unseen candidates. The **Protocol Reader** applies only short screening to this terminal batch, and its predicted inclusions are retained for final output. Recovery performs no reassessment or query-memory update, and no further search round follows.

The final evidence set $\widehat{\mathcal{G}}(p)$ combines the evidence retained during initialization, the final inclusions from all ordinary rounds, and any short-screen inclusions from recovery. This output is accumulated separately from query memory and passed to report synthesis.

4 EXPERIMENTS

4.1 SETUP

Benchmarks. We evaluate medical deep research on **MetaSyn** (Xie et al., 2026) and **CLEF TAR 2018** (Kanoulas et al., 2018). MetaSyn evaluates evidence recovery against its reference inclusion set; CLEF uses abstract-screen relevance. We evaluate both Qwen and GPT on all 86 MetaSyn

reviews and evaluate Qwen on 30 CLEF reviews. Each workflow selects evidence and produces a research report. **Appendix A** describes the data and evaluation protocol.

Baselines and Protocol. We compare seven baselines grouped by evidence collection and synthesis. **Fixed RAG baselines** include BM25-RAG, using BM25 (Robertson & Zaragoza, 2009), Dense-RAG, and MA-RAG. **General-purpose deep research methods** include GPT-R (GPT Researcher) and OpenDR (Open Deep Research), following the MetaSyn comparison (Xie et al., 2026). **Medical deep research methods** include ProtoMA (Xie et al., 2026) and MedSR-Copilot (Huang et al., 2026) in its adapted configuration. **Appendix B.1** details their settings and adaptations.

Implementation Details. MedResearch pairs the frozen MA-Retriever (Xie et al., 2026) with **Qwen3.6-35B-A3B-FP8** or **GPT-5.6 Sol** (`xhigh`) on MetaSyn; CLEF uses Qwen. Local Qwen uses vLLM 0.19.0, while GPT inference is accessed remotely. The same Qwen checkpoint separately evaluates reports from both generation backbones. All model weights remain fixed during execution. We retrieve up to 200 initial candidates, seed query memory from the first 40, and run at most five ordinary search rounds with 10 candidates per batch. The controller can trigger one terminal recovery batch. The default feedback weight is $\lambda = 0.5$. Screening and reassessment expose up to 500 and 2,000 abstract characters, respectively. **Appendix B** gives the full configuration.

Evaluation Metrics. We report seven review-macro metrics in three groups. **Evidence selection** comprises inclusion precision (Inc.P), recall (Inc.R), and F1 (Inc.F1). **Retrieval and retention** comprises candidate recall (Cand.R) and the fraction of retrieved reference evidence retained in the output (Ret.). **Report synthesis** comprises conclusion-direction accuracy (Dir.A) and reference-finding coverage (Insights) (Xie et al., 2026). All main-table scores are percentages. **Appendix C** defines the metrics and aggregation; **Appendix D.1** reports structure quality (SQ) and criteria consistency; **Appendix D.3** reports balanced accuracy, NPV, and candidate F1.

4.2 Main Results

Overall Comparison on MetaSyn. Table 1 shows that **MedResearch achieves the highest evidence-selection recall and F1 under both backbones**, with F1 gains of 14.44 and 9.55 percentage points over the strongest baselines in the Qwen and GPT settings, respectively. It combines broad retrieval coverage with strong retention of eligible evidence. With Qwen, MA-RAG retrieves most reference evidence, yet much of it does not reach its final output. MedResearch achieves comparable candidate coverage while raising retention from 52.64% to 80.54%, turning more of the retrieved evidence into the final evidence base. With GPT, MedResearch combines the highest candidate recall and retention with the second-highest precision, yielding the strongest F1. The advantage extends across fixed RAG, general-purpose deep research, and medical deep research baselines. This pattern aligns with Screen-to-Steer: candidate screening guides continued evidence collection, so the workflow addresses both finding eligible evidence and preserving it for synthesis.

Overall Comparison on CLEF. Table 2 highlights the importance of retaining evidence after retrieval. Dense-RAG achieves the highest candidate recall, but MedResearch obtains higher final recall and F1, improving F1 by 12.80 percentage points. The distinction lies in what happens after candidates are found: MedResearch retains 68.45% of retrieved reference evidence, compared with 43.14% for Dense-RAG, despite their similar candidate coverage. ProtoMA provides a complementary contrast: its relatively strong retention among baselines is accompanied by limited candidate coverage, restricting final evidence recovery. MedResearch combines competitive coverage with the strongest retention, producing the best final evidence-selection F1. These results reinforce the motivation for exclusion reassessment and history-guided collection: retrieving a relevant candidate is only useful when the workflow also recognizes and preserves it as eligible evidence.

Report Synthesis. **MedResearch achieves the highest conclusion-direction accuracy in all three evaluation settings**, extending its strong evidence-selection performance to the final report. Direction accuracy measures alignment with the reference conclusion, while insight coverage measures coverage of reference findings. MedResearch leads both measures on MetaSyn with GPT and ranks second in insight coverage on CLEF, where it also leads in direction accuracy. These results show that the workflow supports both evidence recovery and reporting medically relevant conclusions. This is central to medical deep research: collected evidence must be retained and synthesized into an informative account of the medical question. **Appendix D.1** reports structure quality (SQ) and criteria consistency; **Appendix D.3** reports balanced accuracy, NPV, and candidate F1.

Table 1: **Medical deep research on MetaSyn.** Results with Qwen3.6-35B-A3B-FP8 and GPT-5.6 Sol (`xhigh`); all metrics are higher-is-better.

	Evidence Selection			Retrieval / Retention		Report Synthesis	
Method	Inc.P	Inc.R	Inc.F1	Cand.R	Ret.	Dir.A	Insights
(a) MetaSyn / Qwen3.6-35B-A3B-FP8 86 reviews							
BM25-RAG	55.49	43.96	43.44	74.85	57.03	58.14	30.63
Dense-RAG	**63.74**	48.76	51.01	86.61	55.32	53.49	**34.65**
MA-RAG	63.22	50.01	50.87	91.71	52.64	58.14	34.19
GPT-R	53.74	30.73	35.79	60.75	49.27	47.67	19.30
OpenDR	28.93	28.32	25.91	73.07	36.27	53.49	34.38
ProtoMA	57.40	39.03	41.95	59.21	65.11	46.51	31.09
MedSR-Copilot	59.79	33.86	38.50	83.30	41.91	54.65	24.88
MedResearch	62.53	**75.77**	**65.45**	**93.17**	**80.54**	**65.12**	30.70
(b) MetaSyn / GPT-5.6 Sol (xhigh) 86 reviews							
BM25-RAG	58.81	51.16	48.61	74.55	68.02	40.70	25.88
Dense-RAG	62.82	55.65	51.60	86.45	62.38	45.35	24.71
MA-RAG	55.30	58.77	50.97	91.61	62.49	45.35	26.59
GPT-R	**83.24**	50.13	57.52	75.90	64.17	44.19	26.59
OpenDR	62.51	39.04	41.07	78.64	48.98	41.86	21.41
ProtoMA	67.35	40.95	46.37	67.25	60.56	43.02	25.48
MedSR-Copilot	54.35	26.62	31.38	84.38	31.99	39.53	18.35
MedResearch	74.65	**67.37**	**67.07**	**93.04**	**72.23**	**47.67**	**28.24**

Values are percentages. **Bold**/underline: best/second-best per block, including ties. Metric definitions appear in **Appendix C**.

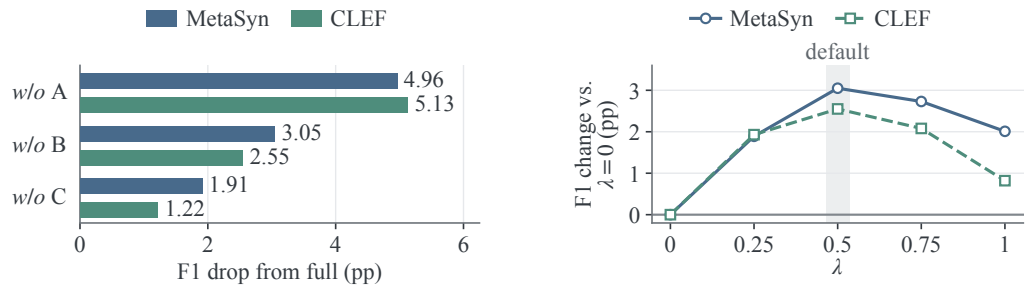
Table 2: **Medical deep research on CLEF with Qwen.** Evidence selection uses the abstract-screen relevance endpoint.

	Evidence Selection			Retrieval / Retention		Report Synthesis	
Method	Inc.P	Inc.R	Inc.F1	Cand.R	Ret.	Dir.A	Insights
CLEF / Qwen 30 reviews							
BM25-RAG	66.03	20.72	27.71	50.19	41.34	46.67	29.33
Dense-RAG	68.13	24.28	31.25	**63.69**	43.14	46.67	34.00
MA-RAG	**71.33**	23.97	30.59	60.65	39.82	33.33	**39.33**
GPT-R	54.57	13.05	18.54	28.73	47.74	50.00	26.00
OpenDR	24.48	8.42	11.09	31.04	22.23	40.00	18.00
ProtoMA	49.76	11.59	16.00	21.47	53.98	43.33	30.00
MedSR-Copilot	67.35	14.70	21.99	55.93	27.21	53.33	35.33
MedResearch	58.26	**42.91**	**44.05**	62.31	**68.45**	**56.67**	38.67

Values are percentages. **Bold**/underline: best/second-best per block, including ties. Metric definitions appear in **Appendix C**.

4.3 Component Ablations

We remove exclusion reassessment (***w/o* A**), Eligibility Memory (***w/o* B**), and the Search Controller (***w/o* C**) individually (Figure 3(a)). **All three removals lower F1 on both datasets.** Removing reassessment causes the largest decline, showing the value of revisiting potential exclusions before they affect retention and feedback. Removing memory weakens the use of prior judgments in later search, while removing the controller eliminates progress-based action selection even when reassessment and memory are retained. Together, the three components revise exclusions, preserve feedback, and control subsequent actions. This coupling turns local screening judgments into trajectory-level search decisions through persistent eligibility feedback across subsequent search rounds. **Appendix E.1** reports all eight configurations.



(a) Component ablations **(b)** Feedback-weight sensitivity

Figure 3: **Component ablations and feedback sensitivity with Qwen.** (a) F1 drop from full system. (b) F1 change from $\lambda = 0$. Values are percentage-point differences on MetaSyn and CLEF.

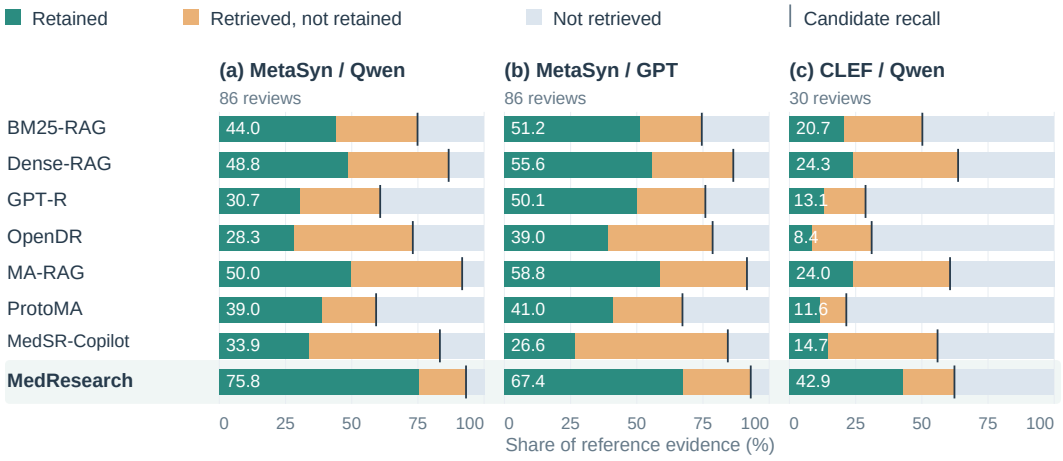


Figure 4: **Where reference evidence is lost.** Bars partition the reference set into retained evidence, retrieved but unretained evidence, and unretrieved evidence. The vertical tick marks candidate recall.

4.4 Hyperparameter Sensitivity

We vary the query-feedback weight λ on MetaSyn and CLEF (Figure 3(b)). At $\lambda = 0$, screening history does not alter the query direction; small weights provide only limited adaptation. **Moderate feedback performs best**: $\lambda = 0.5$ gives the highest F1 on both datasets, with 0.75 close behind. Strong negative feedback can also suppress useful evidence that shares features with excluded records. The decline at $\lambda = 1$ is consistent with giving local screening judgments too much influence over the protocol-derived query. The results favor refining, rather than overwhelming, the original query. **Appendix E.2** provides the five settings.

4.5 Visualization Analysis

Evidence Retrieval and Retention. Figure 4 separates missing evidence into retrieval misses and post-retrieval losses. On MetaSyn with Qwen and on CLEF, MedResearch's retrieved-but-unretained segment is substantially smaller than those of MA-RAG and Dense-RAG, respectively, despite similar candidate coverage. The benefit is therefore visible not only in finding evidence but in preserving what has already been found—the objective of reassessing first-pass exclusions. With GPT, the comparison against GPT-R also shows broader candidate coverage. These distinct patterns connect the final recall gains to both retrieval and retention. This stage-wise view reveals where evidence is lost before report generation.

A Screening-to-Action Example. Figure 5 follows a reference-positive CLEF record. Additional text states that ankle-brachial index (ABI) findings were confirmed by arterial duplex ultrasonography, and the reader revises Exclude to Include. The record is retained for output and enters

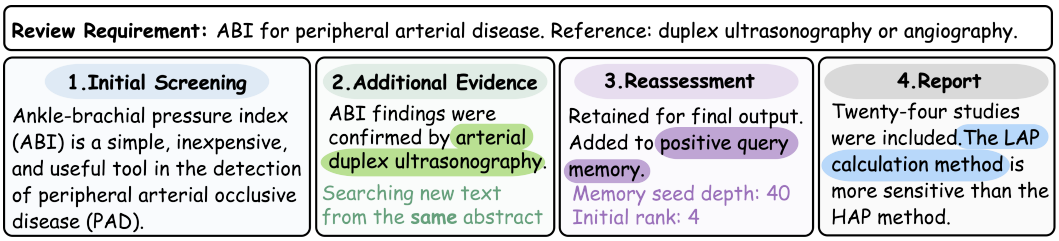


Figure 5: **Screen-to-Steer in a recorded case.** Additional evidence reverses an initial exclusion, updates eligibility memory, and connects the retained evidence to the final report.

positive query memory because it lies within the initial seed. The agent then examines 10 new candidates and stops at zero new inclusions. This sequence shows how a revised screening judgment enters the state used for subsequent actions. It also illustrates how screening decisions persist beyond the current record. **Appendix F** gives the case details.

5 Related Work

General Deep Research Methods. Retrieval-augmented generation (RAG) grounds language models in external evidence through retrieval (Lewis et al., 2020). Building on this paradigm, recent deep research agents perform iterative information seeking and evidence synthesis over multiple retrieval rounds. ReAct (Yao et al., 2022) interleaves reasoning with external actions, while STORM (Shao et al., 2024) organizes multi-perspective retrieval to support long-form report generation. More recent benchmarks, such as DeepResearch Bench (Du et al., 2026), evaluate research agents in terms of both report quality and citation-supported information gathering. These methods demonstrate the effectiveness of agentic retrieval for open-ended research tasks, but they primarily emphasize the quality and factual grounding of the final synthesis. In contrast, medical evidence synthesis additionally requires high-recall evidence collection, where missing protocol-eligible studies may directly affect downstream conclusions.

Medical Deep Research Methods. Medical deep research requires domain-specific retrieval, eligibility screening, evidence extraction, and synthesis under explicit review protocols. Prior work has addressed individual components, including biomedical retrieval with MedCPT (Jin et al., 2023) and automated systematic-review screening with Bio-SIEVE (Robinson et al., 2023). More integrated systems, such as LEADS (Wang et al., 2025b) and TrialMind (Wang et al., 2025a), combine retrieval, eligibility assessment, and structured evidence extraction, while recent agentic frameworks including MetaSyn's ProtoMA (Xie et al., 2026) and MedSR-Copilot (Huang et al., 2026) further support protocol-oriented screening, extraction, meta-analysis, and evidence synthesis. Despite these advances, existing systems largely treat screening as a downstream processing stage rather than persistent feedback for controlling subsequent search. MedResearch instead couples reassessment of first-pass exclusions with persistent eligibility memory, allowing previous screening decisions to guide subsequent retrieval, recovery, and termination. This forms a unified retrieval–screening loop without requiring online human eligibility labels or task-specific retrieval-model updates.

6 Conclusion

We presented **MedResearch**, a training-free medical deep research agent built around **Screen-to-Steer**. By connecting exclusion reassessment, persistent screening memory, and action selection, the agent uses its accumulated judgments to guide evidence collection and retain sources for report synthesis. Experiments on MetaSyn and CLEF show higher evidence-selection recall and F1 than the evaluated baselines, together with strong report-level conclusion alignment. Evidence-flow analysis further shows that retrieval coverage alone is insufficient: retaining eligible evidence through screening is essential to the final synthesis. The retained sources and execution trace also make evidence collection inspectable, linking the report's evidence base to the screening decisions and search actions that produced it. More broadly, the results highlight screening history as a reusable control signal, linking what an agent has learned from evidence to what it does next.

AI USE STATEMENT

The research question and core method were formulated by the authors. We used ChatGPT and Codex to assist with implementing and executing author-designed experiments, providing feedback on implementation and experimental procedures, revising and translating manuscript text, searching and organizing literature, interpreting and presenting results, and preparing LaTeX and scientific figures. The authors made the final methodological and reporting decisions. Separately, language models are components of the evaluated systems, performing evidence screening, exclusion reassessment, and report synthesis. Language-model-assisted evaluation is also used to extract and assess report content against benchmark references, as described in Section 4.1 and Appendix C.4. The authors take responsibility for the final text, code, experimental claims, and visual artifacts, including material prepared with AI assistance.

ETHICS STATEMENT

MedResearch is intended to assist medical evidence research, not to provide clinical advice or replace expert review. Missed eligible evidence, incorrect inclusions, and errors in generated reports can affect downstream conclusions. Reports and selected sources therefore require expert assessment before clinical or policy use. The reported benchmark evaluations concern evidence selection and report assessment; they do not constitute prospective clinical validation.

REPRODUCIBILITY STATEMENT

Section 3 describes the screening, memory, and control procedure. Appendix A details the benchmark settings, Appendix B presents the execution configuration, and Appendix C defines the evaluation measures. Additional quantitative results and the complete component comparison are provided in Appendices D and E.1, respectively. The anonymous repository is linked in the abstract.

REFERENCES

Josien Boetje and Rens van de Schoot. The safe procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic reviews*, 13(1):81, 2024.

Max Callaghan, Finn Müller-Hansen, Melissa Bond, Candyce Hamel, Declan Devane, Wojciech Kusa, Alison O'Mara-Eves, Rene Spijker, Mark Stevenson, Claire Stansfield, et al. Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria. *Systematic Reviews*, 13(1):284, 2024.

Max W Callaghan and Finn Müller-Hansen. Statistical stopping criteria for automated screening in systematic reviews. *Systematic reviews*, 9(1):273, 2020.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deep-research bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026.

Dirk Eyding, Monika Lelgemann, Ulrich Grouven, Martin Härter, Mandy Kromp, Thomas Kaiser, Michaela F Kerekes, Martin Gerken, and Beate Wieseler. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *bmj*, 341, 2010.

Haoming Huang, Qiaoyu Zheng, Pengcheng Qiu, Weike Zhao, Ya Zhang, Weidi Xie, Yanfeng Wang, Xiaoman Zhang, and Chaoyi Wu. A prisma-aligned agentic framework for medical systematic reviews and evidence synthesis. *medRxiv*, pp. 2026–07, 2026.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. CLEF 2018 technologically assisted reviews in empirical medicine overview. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier (eds.), *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL https://ceur-ws.org/Vol-2125/invited_paper_6.pdf.

Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in systematic reviews? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research synthesis methods*, 15(4):616–626, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kentaro Matsui, Tomohiro Utsumi, Yumi Aoki, Taku Maruki, Masahiro Takeshima, and Yoshikazu Takaesu. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using gpt-3.5 and gpt-4 for systematic reviews. *Journal of Medical Internet Research*, 26:e52758, 2024.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

Melissa L Rethlefsen, Shona Kirtley, Siw Waffenschmidt, Ana Patricia Ayala, David Moher, Matthew J Page, and Jonathan B Koffel. Prisma-s: an extension to the prisma statement for reporting literature searches in systematic reviews. *Systematic reviews*, 10(1):39, 2021.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Ambrose Robinson, William Thorne, Ben P Wu, Abdullah Pandor, Munira Essat, Mark Stevenson, and Xingyi Song. Bio-sieve: exploring instruction tuning large language models for systematic review automation. *arXiv preprint arXiv:2308.06610*, 2023.

Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297, 1990.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. Accelerating clinical evidence synthesis with large language models. *NPJ Digital Medicine*, 8(1):509, 2025a.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K Gill, et al. A foundation model for human-ai collaboration in medical literature mining. *Nature communications*, 16(1):8361, 2025b.

Anzhe Xie, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. Benchmarking llm agents on meta-analysis articles from nature portfolio. *arXiv preprint arXiv:2606.17041*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

SUPPLEMENTARY MATERIAL

This supplement covers datasets, implementation, evaluation, complete results, component analysis, and the Reader and report templates.

Contents **Page**

A Datasets and Evaluation Protocol 13
A.1 Benchmarks and Cohorts 13
A.2 MetaSyn: Protocols and a Shared Literature Corpus 13
A.3 CLEF TAR 2018: Review-Specific Screening Pools 13
A.4 Evidence Selection and Report Evaluation 14

B Implementation Details 14
B.1 Baseline Methods and Their Execution Interfaces 14
B.2 Models and Runtime 15
B.3 Execution Configuration 15
B.4 Protocol Reader and Initial Selection 15
B.5 Eligibility Memory and Search Control 15
B.6 Report Synthesis 16

C Evaluation Metrics 16
C.1 Evidence Selection and Alternative Scores 16
C.2 Retrieval and Retention 16
C.3 Candidate-Level Screening Quality 17
C.4 Report Evaluation 17
C.5 Aggregation 17

D Extended Quantitative Results 18
D.1 Complete Report Evaluation 18
D.2 Alternative Evidence-Selection Scores 19
D.3 Candidate-Level Screening Quality 20

E Component Ablations and Sensitivity 21
E.1 Complete Component Matrix 21
E.2 Feedback-Weight Sensitivity 21

F Qualitative Screen-to-Steer Example 22

G Reader and Report Templates 22
G.1 P1: Initial Screening 22
G.2 P2: Exclusion Reassessment 22
G.3 P3: MetaSyn Common Report 23
G.4 P4: CLEF Common Report 23

A Datasets and Evaluation Protocol

A.1 Benchmarks and Cohorts

The evaluation comprises two evidence-discovery settings: a shared scientific literature corpus in MetaSyn and review-specific candidate pools in CLEF TAR 2018. Table A.1 gives the study cohorts. Both MetaSyn backbones are evaluated on 86 reviews; the CLEF evaluation contains 30 reviews.

Table A.1: **Evaluation Cohorts**. Qwen denotes Qwen3.6-35B-A3B-FP8.

Dataset	Generation backbone	Reviews	Evidence-selection endpoint
MetaSyn	Qwen	86	Reference inclusion set
MetaSyn	GPT-5.6 Sol (`xhigh`)	86	Reference inclusion set
CLEF	Qwen	30	Abstract-screen relevance

A.2 MetaSyn: Protocols and a Shared Literature Corpus

MetaSyn is built from 422 expert-curated meta-analysis articles from Nature Portfolio journals. Its construction records research questions, eligibility criteria, published included-study lists, and synthesis targets; these fields are checked against the source articles and supplements. The released split contains 336 training instances and 86 test instances, with a shared corpus of 140,585 PubMed-linked article records (Xie et al., 2026). We use the 86-instance evaluation split with each generation backbone.

Task Inputs and Reference Evidence. The input question is accompanied by Population, Intervention or Exposure, Comparison, and Outcome fields, together with available eligibility and search metadata. The reference set consists of corpus records linked to entries in the source review's included-study list. Thus, recovery is scored at the article-record level. The same article may support more than one review, and one underlying study may have several reports. The public test split contains 1,677 review–article links covering 1,649 distinct linked articles, with 19.5 linked articles per review on average (Xie et al., 2026).

Evidence Access. The released corpus exposes titles, abstracts, publication metadata, and stable record identifiers; some records also have available full-text sections. MedResearch's reading procedure uses the abstract-based contexts described in Appendix B.4. Source-review records are excluded from candidate retrieval, while the linked study reports remain eligible for discovery. The source review supplies the offline reference evidence and report targets.

A.3 CLEF TAR 2018: Review-Specific Screening Pools

CLEF TAR 2018 targets technologically assisted reviews in empirical medicine, with a focus on diagnostic test accuracy. Its new test collection comprises 30 Cochrane reviews. The track distinguishes collection-wide retrieval from screening documents returned by an expert Boolean search (Kanoulas et al., 2018). Our experiments use the latter, review-specific candidate-pool setting.

Each topic supplies a review identifier, research information, and a list of PubMed records. Protocol information describes requirements such as the population, index test, target condition, reference standard, and study design. These fields are mapped to the same question-and-criteria interface used for MetaSyn. Retrieval and screening operate within the topic's frozen pool, preserving the association between each candidate and its record ID.

The evidence-selection endpoint is abstract-screen relevance. In the CLEF annotation scheme, records relevant at this stage include reports that warrant full-text assessment, including some subsequently excluded at full-text review. This differs from final-content inclusion (Kanoulas et al., 2018). The label is therefore used consistently as the reference target for the CLEF precision, recall, and F1 results.

A.4 Evidence Selection and Report Evaluation

For each review, the system receives the question, eligibility criteria, and searchable corpus. It exposes candidates, selects records, and produces a report supported by the selected evidence. Record IDs connect candidate exposure, final selections, and reference membership. The evaluator measures final evidence recovery separately from candidate coverage and report content.

MetaSyn supplies both record-level inclusion targets and review-level synthesis fields. For CLEF, report assessment pairs the topic's evidence-selection labels with review-level criteria and reference findings, extending the screening evaluation to generated synthesis. The same set of eight compared methods is retained in every reported block.

B Implementation Details

B.1 Baseline Methods and Their Execution Interfaces

The baselines span fixed-pool RAG, general-purpose deep research, and medical deep research. Their distinctions concern how candidates are acquired and how eligibility is assessed.

For every method, evidence selection produces a complete, explicit set of selected record IDs before report generation. Evidence-selection metrics are computed directly from these selected-ID sets, independently of the generated report. The selected set is then frozen and passed to the report-generation stage; report text is not used to recover, complete, or modify the selection.

BM25-RAG. The sparse baseline ranks article titles and abstracts using BM25 lexical matching against the protocol-derived query. A fixed ranked pool is then passed to the language model for evidence selection, which returns the complete selected-ID set used for evaluation and subsequent report generation. It provides a lexical retrieval reference for the three fixed-RAG variants (Robertson & Zaragoza, 2009; Xie et al., 2026).

Dense-RAG. This variant substitutes the general-purpose BGE-large-en-v1.5 dense retriever for lexical retrieval. Query and document representations supply semantic similarity scores, while candidate presentation and evidence selection follow the same fixed-pool procedure. The comparison with BM25-RAG therefore examines a different retrieval representation within the same workflow family (Xie et al., 2026).

MA-RAG. MA-RAG uses MetaSyn's MA-Retriever, a BGE-based retriever trained on structured review queries and included-article pairs. It provides a medically relevant, protocol-oriented candidate ranking to the fixed-RAG selector. The retriever is reused without additional training in these runs. MedResearch uses the same retrieval tool within its screening-memory-control loop (Xie et al., 2026).

GPT-R. GPT Researcher builds a research plan, issues queries, gathers source material, and assesses candidate evidence through its multi-step research workflow. Its evidence-selection stage returns the complete selected-ID set before report generation (Xie et al., 2026).

OpenDR. Open Deep Research organizes research through supervisor and researcher stages. Researchers search and inspect evidence, and their findings are combined to determine the selected evidence set. The complete selected-ID set is recorded before report generation, providing a general-purpose multi-step deep-research baseline (Xie et al., 2026).

ProtoMA. ProtoMA is the protocol-oriented pipeline released with MetaSyn. It forms eligibility criteria, generates several PI/ECO-conditioned queries, merges and deduplicates their retrieved candidates, and screens them in batches. Its screening stage produces the selected-ID set used for evaluation and subsequent synthesis (Xie et al., 2026).

MedSR-Copilot. The original framework follows a PRISMA-aligned sequence of retrieval, coarse-to-fine screening, data extraction, risk-of-bias assessment, and synthesis (Huang et al., 2026). For this evaluation, we use its query-generation and evidence-selection stages, including hierarchical screening, context filtering, and final candidate ranking. Available titles and abstracts supply the screening context, and the resulting selected-ID set is fixed before report generation.

B.2 Models and Runtime

The Qwen generator and report evaluator use `Qwen/Qwen3.6-35B-A3B-FP8`. The local Reader uses vLLM with `temperature=0`, `top_p=1`, `n=1`, and `enable_thinking=false`. Output limits are 768 tokens for initial screening and 2,048 for reassessment.

GPT generation uses `gpt-5.6-sol` with `reasoning_effort=xhigh`. Qwen-specific decoding settings apply only to local Qwen calls.

B.3 Execution Configuration

Table B.1 separates candidate acquisition, query-memory initialization, and Reader-call sizes. The initial retrieval contains up to 200 records, all of which are screened. Only the top-ranked 40 records can seed query memory; this prefix does not limit the evidence retained for the final output. Later retrieval operates on unseen records in batches of at most ten.

Table B.1: **MedResearch Execution Configuration**.

Setting	Symbol	Value
Initial retrieval depth	K_0	$\min(200, \lvert\mathcal{D}\rvert)$
Initial memory-seeding depth	s	40
Ordinary / recovery batch size	b	10
Maximum ordinary search rounds	$T_{\max}$	5
Feedback weight	λ	0.5
Initial screening request	—	up to 25 records
Ordinary / recovery request	—	up to 10 records
Initial evidence view	—	first 500 abstract characters
Reassessment context limit	—	up to 2,000 characters

The five-round limit counts ordinary search rounds after initialization. A triggered recovery is one additional terminal batch. Initial Reader calls process at most 25 candidates at a time; search and recovery calls process at most ten. These call-level limits are distinct from the size of the initial candidate pool.

B.4 Protocol Reader and Initial Selection

Reading Contexts. Initial screening presents metadata and the first 500 available abstract characters. Only first-pass EXCLUDE records are reassessed, using the base context and available characters 500–2,000 from the same abstract. With no further text, reassessment still runs with an empty `Additional evidence:` field. It neither requests full text nor supplies inferred content. Initial INCLUDE decisions remain unchanged; reassessed exclusions take their second-pass labels.

For MetaSyn, reassessment uses available abstract characters 500–2,000 when present.

Initial Selection. A predicted initial inclusion is removed when both years are parseable and its publication year exceeds the protocol's search-end year by more than one. Missing or unparseable years leave the decision unchanged, and there is no lower-year bound. The check applies only to initial admission; ordinary search uses final Reader labels and recovery uses short-pass labels.

B.5 Eligibility Memory and Search Control

The initial retrieval order determines the memory-seeding prefix. Retained inclusions in that prefix enter positive memory, and final Reader exclusions enter negative memory. Initial inclusions removed by the year rule enter neither memory. Every initially screened record is marked as visited, including records outside the seed prefix.

In an ordinary round, first-pass exclusions are reassessed before their labels update memory. New final inclusions are retained for output and added to positive memory; final exclusions are added to negative memory. The visited set prevents repeated acquisition, and the new-inclusion count summarizes that round's observed yield. Negative memory steers ordinary queries away from the mean representation of excluded records. Positive memory contributes to the query only in recovery.

The controller starts the first ordinary search when unseen records remain. A productive batch permits another ordinary search until the five-round limit. A zero yield in the first ordinary batch stops collection directly. A later zero can trigger one recovery batch, including at the ordinary-round limit, whereas a productive final ordinary round stops without recovery. Recovery combines positive and negative feedback, short-screens the returned batch, adds its inclusions to the output, and terminates without another query-memory update. Candidate-pool exhaustion also ends collection.

B.6 Report Synthesis

Report generation is separated from evidence selection for all compared methods. Each method first produces a complete selected-ID set, on which all evidence-selection, candidate-recall, and retention metrics are computed. This set is then frozen before synthesis.

For report evaluation, the frozen selected records are supplied to the common report-generation procedure together with the review protocol. The report generator can neither add, remove, nor relabel selected studies, and the generated report is not used to reconstruct the selected-ID set. Consequently, differences in evidence-selection metrics reflect the records selected by each method rather than whether those records are explicitly mentioned in the generated report.

Each selected abstract is limited to 1,200 characters. Reports follow eight sections: Abstract, Introduction, Methods, Results, Key Findings, Discussion, Conclusion, and Limitations, and are instructed to remain under 1,800 words. Appendix F gives a selection-to-report example.

C Evaluation Metrics

C.1 Evidence Selection and Alternative Scores

For a review, let G denote the reference set and L the deduplicated predicted inclusion list after record matching. Define $\mathrm{TP} = |G \cap L|$, $\mathrm{FP} = |L \setminus G|$, and $\mathrm{FN} = |G \setminus L|$. Identifiable output entries are matched to corpus IDs; unmatched entries receive no reference-membership credit and remain in the precision denominator. We use

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

The released inclusion-metric convention assigns zero when a required denominator is zero (Xie et al., 2026).

The supplementary scores change the balance between missed evidence and incorrect inclusions:

$$F_\beta = \frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP} + \beta^2\mathrm{FN} + \mathrm{FP}}, \qquad J = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{C.1}$$

F_2 gives greater weight to recall, $F_{0.5}$ gives greater weight to precision, and Jaccard measures overlap between the predicted and reference sets. They are computed per review before aggregation.

C.2 Retrieval and Retention

Let C be the set of unique records exposed to the workflow and $G_C = G \cap C$. Candidate recall is $|G_C|/|G|$; retention is $|G \cap L|/|G_C|$ whenever the retrieved reference set is nonempty. These quantities describe distinct stages: entering the visible candidate pool and surviving selection into the final evidence set.

The evidence-fate visualization partitions the reference set into retained, retrieved-but-unretained, and unretrieved records. Their per-review shares are $|G \cap L|/|G|$, $|G_C \setminus L|/|G|$, and $|G \setminus C|/|G|$, respectively. This makes post-retrieval omissions visible alongside search misses. The bar segments are averaged shares of the reference set; the separate conditional-retention score uses retrieved reference records as its denominator.

C.3 Candidate-Level Screening Quality

Candidate-level diagnostics use the final inclusion decision restricted to C. Let FP_c be selected nonreference records in C, $\mathrm{FN}_c = |G_C| - \mathrm{TP}$, and $\mathrm{TN}_c = |C| - |G_C| - \mathrm{FP}_c$. The reported diagnostics are

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}_c} + \frac{\mathrm{TN}_c}{\mathrm{TN}_c+\mathrm{FP}_c}\right),$$
$$\mathrm{NPV} = \frac{\mathrm{TN}_c}{\mathrm{TN}_c+\mathrm{FN}_c}, \qquad \mathrm{Cand.F1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}_c+\mathrm{FN}_c}.$$

Balanced accuracy gives equal weight to reference-positive sensitivity and reference-negative specificity. NPV describes the composition of the rejected candidate set. Candidate F1 measures positive-class selection within the exposed pool. Final-output F1 additionally accounts for reference evidence that never entered that pool.

C.4 Report Evaluation

All five report measures are evaluated with the Qwen checkpoint in Appendix B.2, including GPT-generated reports.

Criteria Consistency. The evaluator extracts inclusion/exclusion criteria verbatim from the report and compares them with `inclusion_criteria`/`exclusion_criteria`. Reference criteria are split on semicolons. Matching uses `sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. For normalized reference and extracted criterion embeddings, soft recall averages the maximum cosine similarity for each reference criterion; soft precision reverses the two sets. Their harmonic mean gives Inc.C or Exc.C. Matching is many-to-one rather than a one-to-one assignment.

Conclusion Direction and Findings. The extracted direction must be `Positive`, `Negative`, or `Mixed`; Dir.A measures exact agreement with `Effect_Direction`. Insights compares extracted report statements with `Key_Insights`. The rubric assigns 1.0 for complete coverage, 0.8 for most findings (at least 75%), 0.6 for 50–74%, 0.4 for 25–49%, 0.2 for fewer than 25%, and 0 for no coverage.

Report Structure. SQ uses a 1–5 rubric: 1 indicates missing sections and disorganization; 2 indicates major omissions despite some sections; 3 indicates the main sections are present; 4 indicates complete, ordered sections describing search, selection, and synthesis; 5 additionally includes detailed search, eligibility, quality assessment, and PRISMA-style reporting elements. This is an organization score rather than a clinical-correctness or citation-support assessment.

Reference Construction. MetaSyn uses its released reference fields. CLEF criteria come from the protocol XML; reference findings come from the review PubMed abstract's `main_results` and `authors_conclusions`. Eligibility criteria are task inputs; reference directions and findings are used only for offline assessment.

C.5 Aggregation

Execution cohorts contain 86 reviews for each MetaSyn backbone and 30 for CLEF. Each metric is macro-averaged over its valid evaluation records. Metrics with unavailable reference fields are omitted from that metric's valid set; an empty extracted criterion/insight list receives zero when the applicable reference exists and the evaluator defines that case as zero. For valid direction-reference set $\mathcal{V}$, the accuracy is

$$\mathrm{Dir.A} = \frac{100}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \mathbb{1}[\widehat{c}_i = c_i]. \tag{C.2}$$

The valid direction-reference counts are 86 for MetaSyn/Qwen, 86 for MetaSyn/GPT, and 30 for CLEF. Inclusion-criteria consistency uses 84 valid MetaSyn/Qwen references. Other report fields use their metric-specific valid records. Tables show percentages except SQ, with the best value in bold and the second distinct value underlined, including ties.

D Extended Quantitative Results

The supplementary results cover report quality, alternative evidence-selection scores, and candidate-level screening.

D.1 Complete Report Evaluation

Table D.1 evaluates generated reports along five complementary dimensions. Inclusion- and exclusion-criteria consistency measure agreement with the reference protocol, while direction accuracy and insight coverage assess the conclusions and findings reported. Structure quality captures organization. Together, these measures connect the selected evidence to the resulting synthesis. Definitions appear in Appendix C.4.

Table D.1: **Complete Report-Level Evaluation**.

Method	Inc.C	Exc.C	Dir.A	Insights	SQ
(a) MetaSyn / Qwen3.6-35B-A3B-FP8 86 reviews					
BM25-RAG	57.20	44.21	58.14	30.63	3.6047
Dense-RAG	58.33	44.31	53.49	**34.65**	3.6860
MA-RAG	55.31	44.13	58.14	34.19	3.6395
GPT-R	51.85	45.95	47.67	19.30	3.4302
OpenDR	53.51	48.52	53.49	34.38	3.8140
ProtoMA	54.71	47.22	46.51	31.09	3.8256
MedSR-Copilot	75.91	73.23	54.65	24.88	3.8372
MedResearch	**77.88**	**78.37**	**65.12**	30.70	**3.9070**
(b) MetaSyn / GPT-5.6 Sol (xhigh) 86 reviews					
BM25-RAG	84.23	**87.66**	40.70	25.88	3.1728
Dense-RAG	82.96	86.77	45.35	24.71	3.1529
MA-RAG	84.59	86.43	45.35	26.59	3.1585
GPT-R	83.81	86.76	44.19	26.59	**3.3012**
OpenDR	**84.62**	84.56	41.86	21.41	3.0976
ProtoMA	83.47	86.11	43.02	25.48	3.0723
MedSR-Copilot	83.11	86.92	39.53	18.35	3.1294
MedResearch	84.59	87.18	**47.67**	**28.24**	3.2000
(c) CLEF / Qwen 30 reviews					
BM25-RAG	74.12	61.45	46.67	29.33	3.9333
Dense-RAG	**76.12**	61.59	46.67	34.00	3.9333
MA-RAG	76.09	**63.89**	33.33	**39.33**	3.9333
GPT-R	75.00	60.69	50.00	26.00	**3.9667**
OpenDR	74.57	59.98	40.00	18.00	3.8667
ProtoMA	69.72	56.87	43.33	30.00	3.9333
MedSR-Copilot	72.67	55.86	53.33	35.33	3.8667
MedResearch	74.81	62.13	**56.67**	38.67	3.8667

Inc.C, Exc.C, Dir.A, and Insights are percentages; SQ uses a 1–5 scale. **Bold**/underline: best/second-best per block, including ties.

Protocol Consistency and Synthesis. On MetaSyn with Qwen, MedResearch leads both criteria-consistency measures, direction accuracy, and SQ. With GPT, MedResearch has the highest direction and insight scores and second or joint-second criteria consistency and SQ. On CLEF, it leads direction accuracy and ranks second in insight coverage and exclusion-criteria consistency. These columns complement source recovery with evaluation of the generated report.

D.2 Alternative Evidence-Selection Scores

Table D.2 evaluates the same final selections with recall-weighted F_2, precision-weighted $F_{0.5}$, and Jaccard overlap. This changes the scoring emphasis without changing the selected evidence.

Table D.2: **Alternative Evidence-Selection Scores**.

Method	F_2	$F_{0.5}$	Jaccard
(a) MetaSyn / Qwen3.6-35B-A3B-FP8			
BM25-RAG	42.81	47.27	31.67
Dense-RAG	49.10	<u>55.60</u>	<u>38.32</u>
MA-RAG	<u>49.60</u>	55.02	38.06
GPT-R	32.08	42.86	25.91
OpenDR	26.59	26.90	17.59
ProtoMA	39.52	47.57	29.85
MedSR-Copilot	34.99	45.85	27.14
MedResearch	**70.34**	**62.96**	**51.20**
(b) MetaSyn / GPT-5.6 Sol (xhigh)			
BM25-RAG	49.18	51.14	35.66
Dense-RAG	52.87	53.83	38.70
MA-RAG	<u>54.25</u>	50.94	38.06
GPT-R	52.34	<u>67.06</u>	<u>44.86</u>
OpenDR	38.54	47.56	31.10
ProtoMA	42.20	54.42	34.36
MedSR-Copilot	27.61	39.11	22.00
MedResearch	**66.55**	**69.80**	**54.58**
(c) CLEF / Qwen			
BM25-RAG	22.85	37.97	18.17
Dense-RAG	<u>26.39</u>	<u>41.39</u>	<u>20.67</u>
MA-RAG	25.84	41.34	20.34
GPT-R	14.66	27.25	11.54
OpenDR	9.25	14.86	7.17
ProtoMA	12.85	23.44	9.64
MedSR-Copilot	16.88	33.12	13.69
MedResearch	**42.74**	**48.61**	**30.79**

Values are percentages. **Bold**/<u>underline</u>: best/second-best per block, including ties.

Set-Quality Comparison. MedResearch has the highest F_2, $F_{0.5}$, and Jaccard in all three blocks, extending the F1 comparison to both recall-weighted and precision-weighted scoring and set overlap. On MetaSyn with GPT, the strongest baseline changes from MA-RAG for F_2 to GPT-R for $F_{0.5}$ and Jaccard; MedResearch leads all three views.

D.3 CANDIDATE-LEVEL SCREENING QUALITY

Table D.3 examines selection among the candidates a workflow has actually exposed. It complements final-set evaluation by focusing on whether retrieved reference evidence is retained.

Table D.3: **Balanced Candidate-Level Screening Metrics**.

Method	Balanced Acc.	NPV	Cand.F1
(a) MetaSyn / Qwen3.6-35B-A3B-FP8			
BM25-RAG	<u>77.80</u>	<u>97.05</u>	55.69
Dense-RAG	76.78	96.60	<u>58.20</u>
MA-RAG	75.37	95.73	55.56
GPT-R	70.36	81.06	49.01
OpenDR	65.14	83.52	36.12
ProtoMA	73.19	88.91	56.19
MedSR-Copilot	70.15	94.71	45.07
MedResearch	**87.89**	**98.06**	**67.64**
(b) MetaSyn / GPT-5.6 Sol (xhigh)			
BM25-RAG	<u>83.57</u>	**97.97**	<u>69.77</u>
Dense-RAG	80.75	97.23	66.55
MA-RAG	80.69	96.62	65.45
GPT-R	80.03	89.45	68.62
OpenDR	70.54	88.13	48.41
ProtoMA	75.25	86.16	59.47
MedSR-Copilot	65.40	94.14	35.63
MedResearch	**85.02**	<u>97.72</u>	**69.82**
(c) CLEF / Qwen			
BM25-RAG	<u>68.03</u>	<u>88.57</u>	46.36
Dense-RAG	67.44	84.63	45.30
MA-RAG	66.03	82.16	42.90
GPT-R	64.93	75.95	42.67
OpenDR	58.42	74.03	19.00
ProtoMA	67.21	77.92	<u>47.43</u>
MedSR-Copilot	62.43	80.63	34.30
MedResearch	**77.14**	**88.86**	**58.21**

Values are percentages. **Bold**/<u>underline</u>: best/second-best per block, including ties.

Selection after Exposure. MedResearch leads balanced accuracy and candidate F1 in each block. It also leads NPV on MetaSyn with Qwen and on CLEF, and ranks second with GPT. The positive-class and class-balanced views therefore both show strong retention within the exposed pool, complementing the main evidence-fate decomposition of retrieval and selection losses.

E Component Ablations and Sensitivity

E.1 Complete Component Matrix

Table E.1 reports all eight combinations of exclusion reassessment (A), Eligibility Memory (B), and Search Controller (C). The main figure presents the three removals from the full system; the complete matrix also shows the base and individual additions.

Component Switches. With A off, the 500-character initial screening is retained and no second assessment is made. With A on, first-pass exclusions are reassessed, including when the additional-text field is empty; recovery remains short-only. With B off, ordinary and recovery queries both use q_0, while visited-record exclusion, output accumulation, yield tracking, and recovery eligibility bookkeeping remain active. With B on, ordinary queries use negative memory and recovery uses both positive and negative memory. Initial query memory is seeded only from the first 40 ranked candidates, followed by final ordinary-round judgments. With C off, zero yield does not stop collection: execution proceeds for at most five ordinary batches, or until the pool is exhausted, without recovery. With C on, zero yield ends ordinary search; recovery is permitted only after the second or a later ordinary batch and when a valid positive-memory record exists. Recovery is a single short-screened batch with no memory write-back.

These switches define the component configurations used in the ablation.

Table E.1: **Complete Component Ablations**. Evidence-selection F1 (%).

	Components			F1 (%)	
Configuration	A	B	C	MetaSyn	CLEF
Base	—	—	—	55.10	34.93
A only	✓	—	—	59.51	38.23
B only	—	✓	—	56.10	35.87
C only	—	—	✓	55.19	37.40
A + B	✓	✓	—	<u>63.54</u>	<u>42.83</u>
A + C	✓	—	✓	62.40	41.50
B + C	—	✓	✓	60.49	38.92
Full (A+B+C)	✓	✓	✓	**65.45**	**44.05**

Component Combinations. In the ablation matrix, reassessment is the largest standalone addition, while memory has a larger gain when combined with reassessment than when added alone. The full configuration has the highest F1 in both columns, and each single-component removal lowers it. These comparisons distinguish a component's standalone contribution from its contribution alongside reassessment, query feedback, and action selection.

E.2 Feedback-Weight Sensitivity

Table E.2 provides the raw F1 values for the five feedback weights. The compact main-text plot expresses the same values relative to $\lambda = 0$.

Table E.2: **Sensitivity to the Feedback Weight**. Evidence-selection F1 (%) with Qwen.

λ	**MetaSyn**	**CLEF**
0	62.40	41.50
0.25	64.29	43.43
0.5 (default)	**65.45**	**44.05**
0.75	<u>65.13</u>	<u>43.58</u>
1	64.41	42.32

At $\lambda = 0$, screening judgments do not alter query direction. The default 0.5 is best and 0.75 is second in both datasets. The decline at weight 1 is consistent with overly strong feedback: negative memory can suppress useful candidates that share semantic features with excluded records. The tested values favor moderate correction of the protocol-derived query.

F QUALITATIVE SCREEN-TO-STEER EXAMPLE

Figure 5 follows CLEF review `CD010680` on ankle-brachial index (ABI) and record `16950430`. The protocol requires duplex ultrasonography or angiography as the reference standard. Initial screening returns EXCLUDE; additional abstract text reports confirmation by arterial duplex ultrasonography, and reassessment returns INCLUDE.

The reassessed record is retained and contributes to the final evidence set. The resulting report reflects the selected studies and illustrates the end-to-end trajectory from evidence reassessment to synthesis.

G READER AND REPORT TEMPLATES

Common report input. The following report templates are applied after evidence selection to all compared methods. For each method, the input consists of its complete frozen selected-ID set and the corresponding study records. The report generator does not perform additional evidence selection, and its output is not used to define or recover the selected-ID set.

G.1 P1: INITIAL SCREENING

```
Screen each candidate against the systematic-review protocol.
Return exactly one include or exclude decision for every candidate index.
Use only the title and abstract shown; do not infer hidden fields or
↪ labels.
{protocol}

Candidates:
Candidate index: {index}
Corpus ID: {corpus_id}
Title: {title}
Publication year: {year}
Abstract: {initial_abstract_context}
```

G.2 P2: EXCLUSION REASSESSMENT

```
Make one final binary eligibility decision for every candidate.
Apply the protocol literally and use only the visible base and additional
↪ evidence.
Return include only when the visible evidence supports eligibility; do
↪ not invent missing facts.
{protocol}

Candidates:
Candidate index: {index}
Title: {title}
Publication year: {year}
Base evidence: {base_evidence}
Additional evidence: {additional_evidence}
```

G.3 P3: METASYN COMMON REPORT

> Write a concise systematic-review report using only the protocol and included study records below. The evidence selection is frozen by the caller: do not add, remove, or relabel studies. Use exactly these top-level headings in this order: Abstract, Introduction, Methods, Results, Key Findings, Discussion, Conclusion, Limitations. Under Methods, explicitly cover the review question, date-bounded search, verbatim inclusion and exclusion criteria, study selection counts, data extraction, quality-assessment limitations, and qualitative evidence synthesis. Do not invent databases, PRISMA counts, risk-of-bias assessments, effect sizes, or statistics. Keep the report under 1,800 words. Mention at most 12 representative provided corpus IDs, and never invent IDs or extrapolate ID ranges. State conclusion direction explicitly as exactly one of Positive, Negative, or Mixed. For studies without an available abstract, do not infer study findings.

```
PROTOCOL
{protocol}

FROZEN COUNTS
retrieved_count: {retrieved_count}
included_count: {included_count}

INCLUDED STUDIES
{included_studies}
```

G.4 P4: CLEF COMMON REPORT

> Write a concise systematic-review report using only the protocol and included study records below. The evidence selection is frozen by the caller: do not add, remove, or relabel studies. Use exactly these top-level headings in this order: Abstract, Introduction, Methods, Results, Key Findings, Discussion, Conclusion, Limitations. Under Methods, explicitly cover the review question, verbatim selection criteria, study selection counts, data extraction, quality-assessment limitations, and qualitative evidence synthesis. Do not invent databases, PRISMA counts, risk-of-bias assessments, effect sizes, or statistics. Keep the report under 1,800 words. Mention at most 12 representative provided candidate IDs, and never invent IDs. State conclusion direction explicitly as exactly one of Positive, Negative, or Mixed. For studies without an available abstract, do not infer study findings.

```
PROTOCOL
review_title: {review_title}
background: {background}
objectives: {objectives}
selection_criteria: {selection_criteria}

FROZEN COUNTS
included_count: {included_count}

INCLUDED STUDIES
{included_studies}
```